Statistical model specification and power: recommendations on the use of test-qualified pooling in analysis of experimental data

Nick Colegrave[1] and Graeme D Ruxton[2]

1. School of Biological Science, University of Edinburgh, Edinburgh EH14 4AJ, UK

2. School of Biology, University of St Andrews, St Andrews KY16 9TH, UK

Abstract

A common approach to the analysis of experimental data across much of the biological sciences is test-qualified pooling. Here non-significant terms are dropped from a statistical model, effectively pooling the variation associated with each removed term with the error term used to test hypotheses (or estimate effect sizes). This pooling is only carried out if statistical testing on the basis of applying that data to a previous more complicated model provides motivation for this model-simplification; hence the pooling is test-qualified. In pooling, the researcher increases the degrees of freedom of the error term with the aim of increasing statistical power to test their hypotheses of interest. Despite this approach being widely adopted and explicitly recommended by some of the most widely-cited statistical textbooks aimed at biologists, here we argue that (except in highly specialised circumstances that we can identify) the hoped-for improvement in statistical power will be small or non-existent, and there is likely to be much reduced reliability of the statistical procedures through deviation of type I error rates from nominal levels. We thus call for greatly reduced use of test-qualified pooling across experimental biology, more careful justification of any use that continues, and a different philosophy for initial selection of statistical models in the light of this change in procedure.

Key words: experimental design, pseudoreplication, model simplification

Introduction

A common approach to the analysis of experimental data across disparate parts of the biological sciences is test-qualified pooling. A common manifestation of this approach can be summarised as follows: the researcher fits their data to a model that they select on the basis of the design of their study and the hypotheses they are interested in testing. After examining the significance of terms in the model that are not specifically related to the hypothesis currently under investigation, the researcher then removes non-significant terms from the model, and re-fits their data to this simplified model. That is, some terms were included in the original model not because they allow an interesting hypothesis to be tested but because (on the basis of the specifics of the experimental design allied to previous knowledge of the system) they were expected to explain substantial portions of the variation. If the data generated in this particular experiment do not suggest that one or more of these terms are strongly influential then they are dropped from the model, and further analysis is performed based on a simplified model. Such a simplification process is often seen as attractive in making presentation of results more compact, in highlighting more influential variables, and/or in increasing statistical power for exploring the significance of remaining terms. By simplifying the model in this way, the researcher is effectively *pooling* the variation associated with each removed term with the error term that will ultimately be used to test their hypotheses. This pooling is only carried out if statistical testing on the basis of applying that data to a previous more complicated model provides motivation for this approach, hence the pooling is *test-qualified*. In pooling, the researcher increases the degrees of freedom of the error term with the aim of increasing statistical power to test their hypotheses of interest. Despite this approach being widely adopted and explicitly recommended by some works on data analysis (e.g. [1]), other influential authors explicitly warned against this practice (e.g. [2]). Here we want to offer some resolution of this apparent conflict in the literature, in order to help authors, reviewers, editors and readers evaluate the consequences of pooling in different circumstances. Note that although we couch this discussion in terms of null-hypothesis statistical testing, the arguments transfer naturally to

approaches based on estimation of effect size; our discussion is however focussed on the analysis of data from planned experiments rather than from purely observational studies. The costs and benefits of test-qualified pooling are more clear-cut for planned experiments where potential confounding factors can often be eliminated or controlled for by careful experimental design, removing the need to deal with these factors statistically. Also, planned experiments generally are of what is termed a "confirmatory" nature, where the study specifically aims to test one or more hypotheses known from the outset. Observational studies more often have an "exploratory" motivation involving measuring a broad range of variables and then seeking to rank them in terms of potential importance and influence. We return to these issues in the *Discussion*.

Being clear what pooling is and why you might want to do it

To clarify the issues we consider a specific example. You are interested in the effect of an experimental treatment (a new humidification system) on the growth of individually-potted tomato plants. Your experiment will be conducted in ten small greenhouses at your research station, and the nature of the treatment means that it has to be applied to whole greenhouses. You install the humidification system in five (randomly selected) greenhouses, leaving the other five as controls, and you assay the growth of 40 tomato plants in each greenhouse. In this design the greenhouse is the experimental unit, and any hypothesis test of the treatment should use an error based on the variation amongst greenhouses rather than variation amongst the individual plants. In this case the simplest means of analysis would be to calculate a mean growth rate across the 40 plants in each greenhouse and carry out a one-way ANOVA using these 10 independent data points.

However, as a thought experiment, suppose that we somehow knew for a fact that growth conditions (in the absence of our treatment manipulation) were absolutely identical amongst our greenhouses. In this imaginary situation we might argue that, since greenhouse-to-greenhouse variation is not confounded with any treatment effect we can use the growth measures from the individual plants as independent data points in our analysis. This will result in a substantial increase

in our degree of freedom, and consequently our statistical power to detect treatment effects. Of course in reality, we cannot usually know with certainty whether our greenhouses vary, and this has led to the development of methods for test-qualified pooling. In this case, we would start by fitting the nested model defined by the design of our study (with individual plants being nested within greenhouse). This would include the treatment term, a nested term for the variation amongst greenhouses in the same treatment group, and a second error term corresponding to the variation amongst plants in the same greenhouse. The key to test-qualified pooling is that the set of data itself influences the nature of the analyses performed on it. If initial analysis of the full model indicates substantial variation amongst greenhouses, then the significance of the treatment term is tested using the variation amongst greenhouses as its error term with 8 df. However, if there is no evidence of substantial greenhouse-to-greenhouse variation in this initial analysis then the among-greenhouse and the true error variations are pooled, and this combined error term with 398 df is used then to provide a test of the treatment effect that is expected to benefit from higher statistical power (see [3-5] for commonly-cited texts that recommend this approach). The justification that advocates of test-qualified pooling give for this approach is that in the absence of any greenhouse effect, the among-greenhouse and the within-greenhouse error terms are both estimating the same thing, and so by combining them we get a better estimate than we would estimating the two separately.

However pooling is not limited to nested designs. Continuing with tomatoes and greenhouses, you now want to compare the effects of four different growing media in individually-potted tomato plants rather than the effect of humidity. To gain a sufficient sample size for the experiment you have to use three different greenhouses to keep all the plants, but because your treatments can now be applied randomly to individual plants, you randomly allocate equal numbers of plants to each treatment in each greenhouse leading to a randomised block design (with specific greenhouse identity as the blocking factor, with three levels). The statistical model implied by this design would include terms for both treatment applied to a plant and the specific greenhouse a plant was kept in,

as well as a treatment-by-greenhouse interaction and an amongst-plant error term based on the variation amongst individual plants within the same treatment-greenhouse combination. Depending on the exact hypothesis we wish to test, the appropriate error term for our treatment effect will be either the interaction term, or the amongst-plant error term [6], but in either case, if the interaction term is not significant, we might chose to pool its variation with the amongst-plant error term prior to testing the treatment effect. Similarly, we might then decide that if the greenhouse term is also non-significant, we would add that source of variation and its associated degrees of freedom to our error pool. In either case, we would be carrying out test-qualified pooling.

Another form of pooling can involve the initial test that triggers whether pooling is used or not being entirely separate to the model testing the hypotheses of interest. To illustrate this, we return to the experiment above comparing the effects of four different growing media on individually-potted tomato plants. Imagine that, because of a change of supplier at your institute, you ended up using two different but broadly similar types of pots to grow the tomatoes in. Plants are randomised to pot type as well as to growth medium and greenhouse. You really do not expect type of pot to influence growth rates, but just to be careful you first of all perform a t-test comparing growth rates across the two types of pot. Your plan is that if (as you expect) this t-test reveals no evidence of a difference, you report this and use this test as justification for pooling data across the two pot types in your subsequent analyses. However if it does reveal evidence of a difference then you will either add pot-type as a factor in subsequent analyses or carry out separate analyses for the two types of pot. Again, there is the potential for pooling driven by the results of a pre-test, so this scenario is another manifestation of test-qualified pooling.

Why is test-qualified pooling controversial?

The case against pooling was made most forcefully and explicitly in the biological literature by Stuart Hurlbert primarily in relation to its use in nested designs [2]. Hurlbert coined the expression *pseudoreplication* for the situation where authors treat data-points that are not independent as if

they were independent in their data analysis. His original paper on this [7] has been cited over 6000 times and has been hugely influential in the design of data collection and the analysis of data spanning all of biology. Hurlbert considers the pooling of errors in a nested analysis to be a form of pseudoreplication, a form that he calls *test-qualified sacrificial pseudoreplication*. He argues that pooling biases p-values downwards and biases confidence intervals towards being too narrow. He further argues that demanding a higher p-value than 0.05 in the initial test before pooling (a process often called "sometimes pooling") reduces but does not eliminate these problems. An analogous argument can be made against pooling interaction terms with error terms when analysing randomised block designs [6]. However, even in situations where pooling might not be regarded as analogous to pseudoreplication (e.g. pooling an interaction between two fixed factors prior to testing the main effects), type 1 error rates can be increased (as we will see below). Despite this, pooling is still regularly practiced, and is recommended in influential statistics textbooks aimed at biologists (e.g. [3-5]) and research papers on statistical methodology (e.g. [2,8]). In the next section we argue that both philosophically and pragmatically there are strong arguments for siding with Hurlbert.

The philosophy and pragmatics of pooling

The two main philosophical arguments against pooling are well articulated by Newman et al. [7], and can be explained in the context of our greenhouses and growth media example. Firstly, if we use pooling, then the way that we test for an effect of growth medium becomes conditional on the data, but that conditionality is not acknowledged in the associated p-values. That is, whether we test the effect of medium in a model with or without a *greenhouse* term will be determined by the data. Philosophically, p-values are probabilities based on a very large number of notional replicates of exactly the experiment under investigation. So imagine that we repeat the full experiment and analysis of the resulting data again and again. In replicates of this experiment, if we adopt a test-qualified pooling approach then sometimes the analysis will test the main hypothesis one way and

sometimes the other. For each form of the analysis, that particular analysis will be implemented only for a specific subset of replicate experiments determined by the patterns of data in that replicate experiment. Importantly, this is a biased sample of all the possible replicate experiments in terms of properties of the sample. Yet the test is predicated on the assumption that it is applied to data from an experiment drawn without bias from the population of all possible replicates of this experiment. It is this mismatch that leads to lack of control of type I error and of confidence intervals. Secondly, by pooling (no matter what critical value we compare the calculated p-value against) we are accepting that the null hypothesis that there is no effect of *greenhouse* is true, and the whole philosophy of null-hypothesis statistical testing is that the null hypothesis is never accepted as true, rather we might either reject it or find that we do not have sufficient grounds to reject it. Thus, from a purist philosophical perspective pooling should not be recommended.

We next ask if there is a pragmatic argument that says that pooling may have some less-than-ideal properties, but pooling leads to relatively mild misbehaviours that are sometimes outweighed by the (enhanced power) benefits of pooling. There is no underlying theory to give general and definitive answers to the issue of pragmatics raised above; all we have to go on are a number of numerical explorations of specific cases. However, the consensus in this literature is that (i) pooling can cause actual type one error rates to be very different from the nominal value, and (ii) there is no consistent and substantial increase in power to compensate. Walde-Tsadik & Afifi [9] explore the effect of always pooling when one factor is associated with a p-value above 0.05, and also of "sometimes pooling" when the required critical value was higher than 0.05 in two-way ANOVA random effects models. They found that both procedures very rarely offered adequate control of type-1 error rate and even less commonly lead to significant improvement in power to test for an effect of the other factor. Hines [10] performed extensive simulations and concluded that for multifactorial ANOVA "the conditions for pooling to be even potentially rewarding are more restrictive than might be expected, and power improvements are generally lower". Janky [11] performed a similar analysis of split-plot designs and concluded that "pooling generally inflates Type I error and offers at best

insubstantial gain in power (and often power loss) relative to the nominal test." Even when using a conservative "sometimes pooling" value of $\alpha = 0.35$ to trigger pooling, Janky found the type I error rate in subsequent tests on pooled data rose from the nominal 5% to generally somewhere between 7% and 11%. This study was interesting for highlighting that pooling actually led to a reduction of power more often than it lead to a substantial gain in power; this occurs because the increase in inherent variation caused by pooling dominates any effect of increased degrees of freedom devoted to exploring remaining factors. Figure 1 shows examples of deviations in both directions from the nominal 5% level for type I error rates generated by simulations of our whole-greenhouse-treatment thought experiment. In exploring our model we found that small changes in parameter values could lead to substantial change in the magnitude and direction of deviations from the nominal level. It is difficult to make generalisations about the circumstances under which deviations will be strongest. In common with the other studies discussed directly above, we found that the direction and magnitude of deviations are driven by a complex interaction between structure of the experimental design, aspects of the shape of the underlying "population" from which sample values are obtained, and sample sizes. Also, as the highest line in Figure 1 illustrates, relationships with parameter values can be non-monotonic.

Discussion and Conclusion

Use of test-qualified pooling is widely adopted, but its prevalence across biological sciences is patchy. For example, it is much less commonplace in clinical trials; where often statistical analyses have to be specified in pre-registration of trials, and thus scope for flexibility in data analysis is reduced. Test-qualified pooling is also relatively uncommon in the agricultural sciences, where particular designs and modes of analysis that avoid issues of pooling are traditional; and the statistical software package *Genstat* is commonly used, which is particularly suited to forms of analyses that avoid test-qualified pooling.

We do still consider that test-qualified pooling is over-used in biology. Simply, in "confirmatory studies" based on designed experiments where we aim to test specific hypotheses (or estimate specific effect sizes) we do not recommend pooling under any circumstances. The often-modest expected increases in power from pooling do not make it an attractive option when its drawbacks are taken into account. Apart from statistical power, the other attraction to pooling is simplification of the presentation of results, but we feel that this will never be sufficient grounds for justifying the process. We would only recommend pooling in such a study if the decision to consider test-qualified pooling was made on the basis of a prior simulation study that aimed at evaluating the consequences of pooling for Type I and Type II error rates. We have yet to see an example of a study that provided such a justification for pooling.

As we mentioned in the *Introduction*, it is not as easy to offer clear and simple guidance on pooling in purely observational studies, and studies where the researchers' aims are more focussed on exploration or prediction than on testing specific hypotheses. However, in such situations pooling can be seen as a facet of *model selection* – which is an area of considerable activity in applied statistics. A particularly useful introduction to the concepts involved is that of Chatfield [12]. He makes the point that if the same data-set is used to both select the most appropriate model from a suite of alternatives and also to fit that model, then the interpretation of the fitted model should be quite different from circumstances where the form of the model is decided upon first and only then is the data applied to fit that model. Where there is uncertainty as to the most appropriate model, then there are methodological developments in *model averaging* that can acknowledge this ([13] and [14] offer good introductions for the biologist). A failure to properly acknowledge model uncertainty when the same data is used to select and fit the model can read to very unreliable inferences ([12],[15],[16]).

Despite the complexity of the literature on model selection and model uncertainty, we feel that we can offer a general opinion on the utility of test-qualified pooling outside designed experiments. For

more exploratory studies where the intention is to identify factors that might be of interest, rather than to test specific hypotheses, then test-qualified pooling might be more attractive; since researchers may be willing to live with loss of control of type I error rates if this helps boost their statistical power to flag up factors of interest. That is, they may be prepared to suffer higher rates of false positives to boost their likelihood of detecting real effects. We expect that these power gains may sometimes be considerable for nested-designs. However for other types of design the literature discussed in the last section should serve as a caution that power gains from pooling may be small or non-existent. Our view is that even in exploratory studies, test-qualified pooling cannot really be recommended except perhaps where the design is nested and where the size of the experiment was reduced from its ideal size by practical constraints or unforeseen adverse circumstances.

Where does this leave the experimenter in our tomato plant example who just wanted to be diligent and reassure themselves and their readers that there was no effect due to two different types of pots being used? They have to make a decision about how important this check is to them. If they feel that it is worth investing a few degrees of freedom in, then they should include *type-of-pot* as a factor in their analysis and pay a modest cost in reduced power to test the hypothesis (comparing different growth media) that they are really interested in. Alternatively, they may decide that careful experimental design and explanation of that experimental design should allay concerns about differential effects of pot types sufficiently that there is no need for formal statistical testing. More generally, we all have to accept that there are no free statistical analyses, and think hard about which factors to include in any model. This is analogous to the decision to block on a given variable in experimental design. It is only advantageous to block on variables that explain a substantial fraction of variation between experimental units, otherwise the degrees of freedom lost in including that blocking term are not compensated for by effective partitioning of variation into error and other terms.

Sometimes we can make a strong enough case based on careful experimental design (especially use of randomisation), biological intuition, and logical reasoning for why we can safely assume that some potentially influential factors are in fact very unlikely to be important in our study, and so we omit them from our statistical procedures. In fact, we do this all the time. In our example the researcher felt no need to test whether which shelf on a greenhouse a pot was placed on had an effect, or what side of the greenhouse, or how near to the door of the greenhouse it was. Sometimes we will feel that we cannot make a sufficiently strong case this way, and we should then include that factor in our model and explore its effects statistically. As so often in the design and analysis of scientific studies, there are no black-and-white rules for which factors to include in your statistical model; we need to think hard about it and justify our choices in terms of experimental design, understanding of underlying biology and logical reasoning. This should be good news: model selection should be much more about biology than about mathematics and probability theory – and biology is what we are interested in.

Acknowledgment: We thank Gavin Gibson and three anonymous reviewers for perceptive comments.

Author contributions: This article was conceived, developed and written equally by both authors.

References

[1] Schank, J. C., & Koehnle, T. J. (2009). Pseudoreplication is a pseudoproblem.*Journal of Comparative Psychology*, *123*(4), 421.

[2] Hurlbert, S. H. (2009). The ancient black art and transdisciplinary extent of pseudoreplication. *Journal of Comparative Psychology*, *123*(4), 434.

[3] Sokal, R. R., & Rohlf, F. J. (1995). Biometry: the principals and practice of statistics in biological research. *WH Freeman and Company, New York*.

[4] Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

[5] Zar, J. H. (1999). *Biostatistical analysis*. Pearson Education India.

[6] Newman, J. A., Bergelson, J., & Grafen, A. (1997). Blocking factors and hypothesis tests in ecology: is your statistics text wrong?. *Ecology*, *78*(5), 1312-1320.

[7] Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological monographs*, *54*(2), 187-211.

[8] Crits-Christoph, P., Tu, X., & Gallop, R. (2003). Therapists as fixed versus random effects-some statistical and conceptual issues: a comment on Siemer and Joormann (2003).Psychological Methods 8, 518-523.

[9] Wolde-Tsadik, G., & Afifi, A. A. (1980). A comparison of the “sometimes pool”,“sometimes switch” and “never pool” procedures in the two-way ANOVA random effects model. *Technometrics*, *22*(3), 367-373.

[10] Hines, W. G. S. (1996). Pragmatics of pooling in ANOVA tables. *The American Statistician*, *50*(2), 127-139.

[11] Janky, D. G. (2000). Sometimes pooling for analysis of variance hypothesis tests: A review and study of a split-plot model. *The American Statistician*,*54*(4), 269-279.

[12] Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A, 158, 419–466.*

[13] Whittingham, M. J., Stephens, P. A., Bradbury, R. B. & Freckleton, R. P., (2006). Why do we still use stepwise modelling in ecology and behaviour?. *Journal of Animal Ecology*, *75*,.1182-1189.

[14] Richards, S. A., Whittingham, M. J. & Stephens, P. A., (2011). Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behavioral Ecology and Sociobiology*, *65*, 77-89.

[15] Blanchet, F. G., Legendre, P. & Borcard, D., (2008). Forward selection of explanatory variables. *Ecology*, 89, 2623-2632.

[16] Mundry, R. & Nunn, C. L., (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. *The American Naturalist*, *173*, 119-123.

Figure legend

Figure 1: To illustrate how the type 1 error rate can be affected by test qualified pooling we examined simulated data sets for both 4 (broken line) and 10 (solid line) greenhouses. In both cases, equal numbers of greenhouses were allocated to control or treatment conditions (but condition had no effect on plant growth), and 40 plants were measured in each greenhouse. We also examined the effect of two different alpha levels for the pooling decision (recommended in [3]: open circles = 0.25 and closed circles = 0.75), and several different levels of among-greenhouse variation (σ^2). Under many different parameter combinations the actual type 1 error rate differs from the desired value of 0.05, sometimes substantially.

Plant growth rates were calculated as a baseline value (10) plus an individual deviation drawn from N(0,1) plus a greenhouse-deviation drawn from N(0,σ) and the same for all plants in a given greenhouse. We analysed each data set in two ways. First we carried out a nested analysis of variance in which the treatment mean square was tested over the among-greenhouses within-treatment mean square. The same analysis tested for variance among greenhouses by comparing the among-greenhouses mean square to the amongst-plants error mean square. Second we carried out an analysis in which data from all greenhouses was pooled. The decision as to which P value to use for our actual hypothesis test for the effect of the treatment was based on the significance of the among-greenhouse test at one of two alpha levels. If this test was significant at the appropriate alpha level we used the P value from the nested model, otherwise we used the P value from the second model. This process was repeated 100000. The proportion of these runs that gave a P value

of less than 0.05 (i.e. a false positive at alpha = 0.05) is an estimate of the type 1 error rate. The simulations were carried out in R, with the AOV function being used for the analyses.

Figure 1: